UNIVERSITY OF CANTERBURY
Te Whare Wānanga o Waitaha
CHRISTCHURCH NEW ZEALAND

Time Series and Linear Regression

Statistics and Modelling Scholarship Workshop

August 21, 2010

Don McNickle
364-2666
don.mcnickle@canterbury.ac.nz

Management Science Group
Department of Management
University of Canterbury

Management Science is part of the Management Department at Canterbury
http://www.mang.canterbury.ac.nz/

For information on recent student Honours Projects in Management Science go to:
http://www.mang.canterbury.ac.nz/courses/msci680.shtml

Time Series

Beliefs:

Components:

Many time series contain a number of **predictable components** (**trend, seasonality, cycles)** together with some uncontrollable **erratic variation**.

We **decompose** the series and forecast the value based on the predictable components.

Continuity:

We believe that patterns in the data will continue some time into the future

Trend

An upward or downward tendency in the data.

Average level of the series changes over time.

Long term movement, predominantly in one direction only.

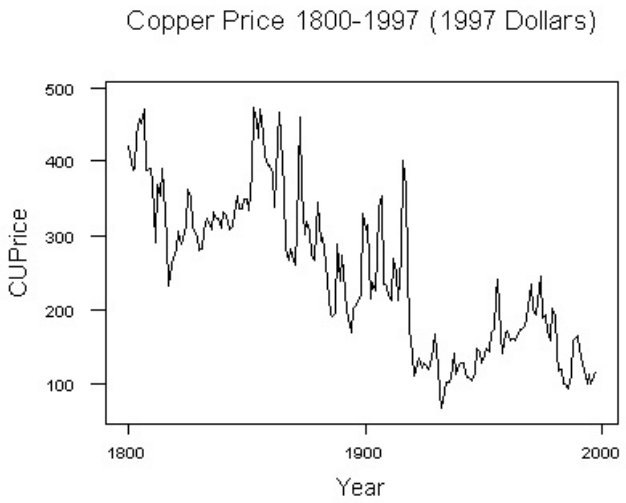
Caused by:

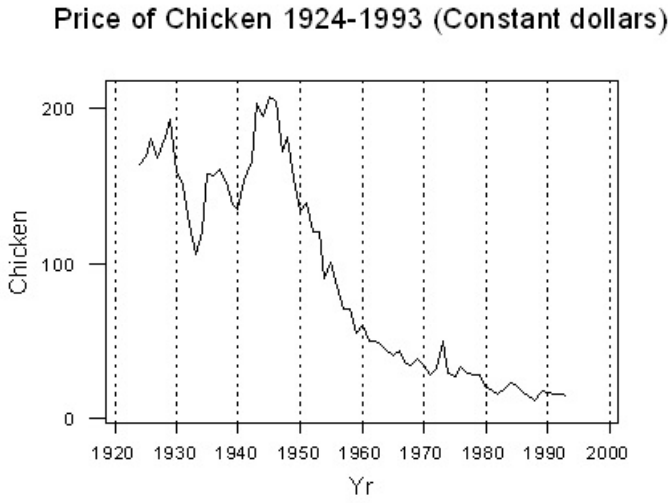
Growth in a company, markets.
(Predictable components e.g inflation, increasing population should be removed.)

Increase in price or taxation, change in habits of the population.

On the average, does the price of commodities, e.g. a pound of copper metal, a pound of chicken, increase or decrease with time?

Series with Trend





Read http://en.wikipedia.org/wiki/Simon-Ehrlich_wager for more on this

Seasonality

- A regular pattern of fluctuations that repeats from year to year (week to week). - calendar dependent

- Often one of the "valuable" components that leads to forecasting success.

Reasons:

Calendar related factors:

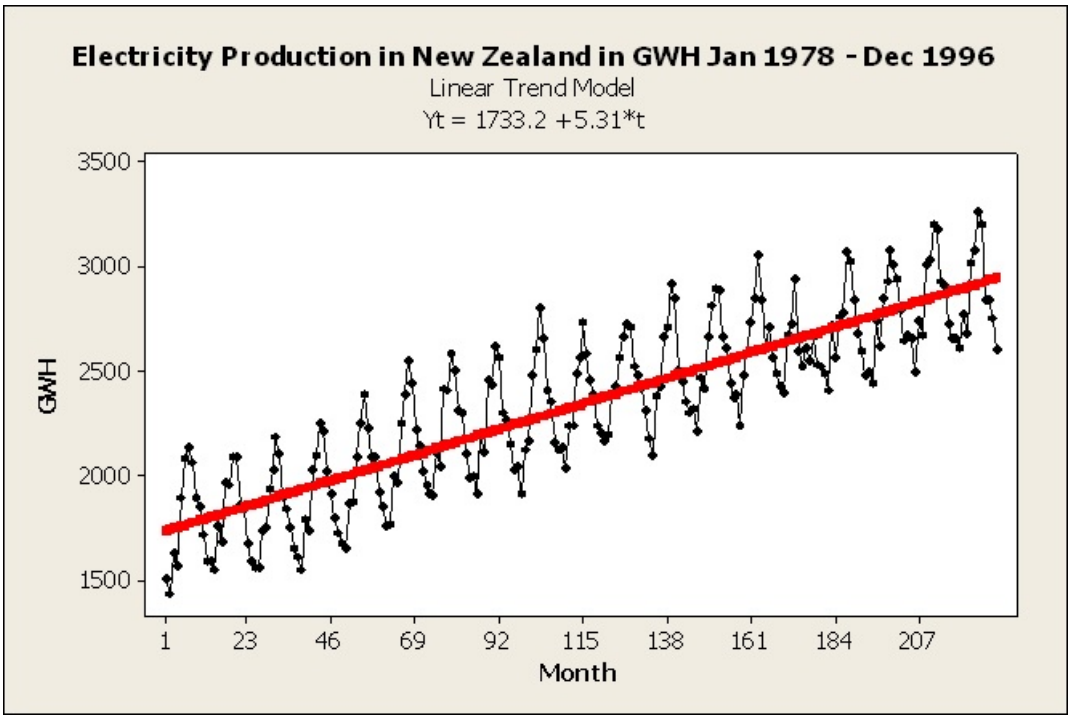
Holidays, Christmas, School terms etc.

Weather and temperature related factors that follow on from the calendar:

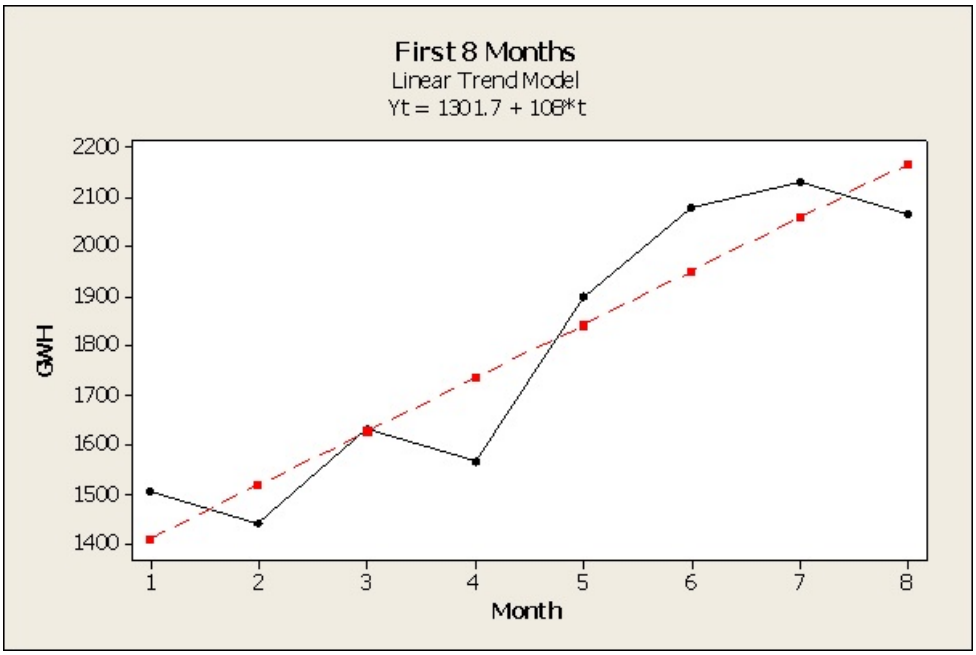
electricity consumption, food production, growth patterns, etc.

However, we can't directly forecast a series that has a seasonal component. Why?

I need a nice regular series here:

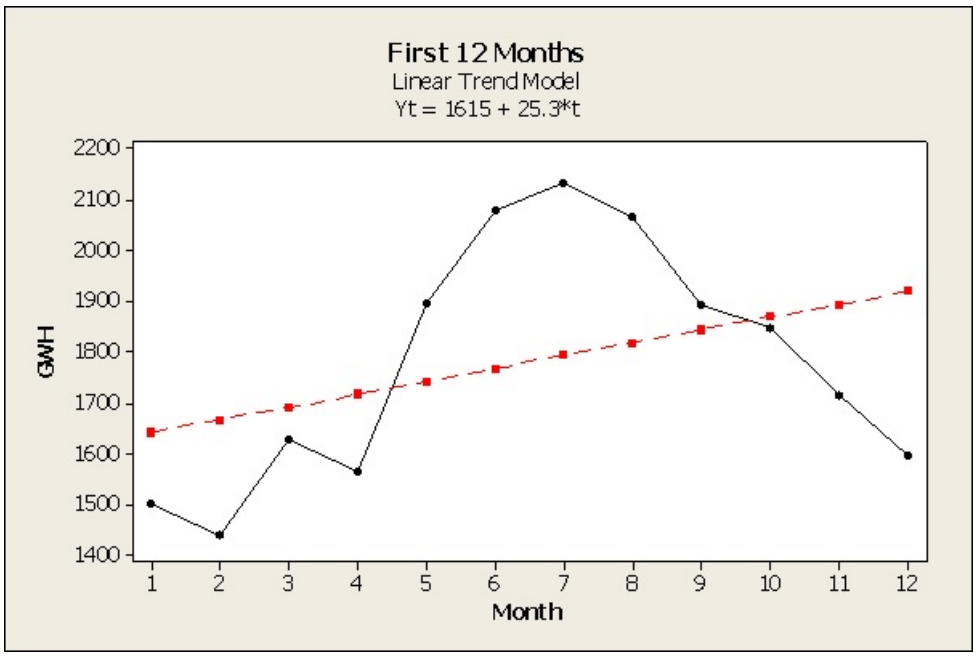


What if we were just starting out and had only eight months of data?



Then we come back after 10 more months.....

The trend line wobbles with the additional seasons.



We need to (temporarily) get rid of seasonality, to produce **deseasonalised (seasonally adjusted) data.**

How to get rid of seasonality temporarily?

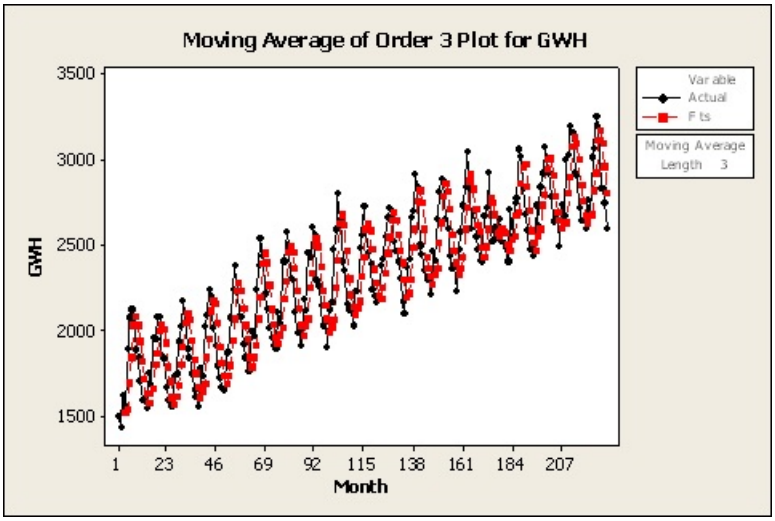
Moving Averages (Moving Means)

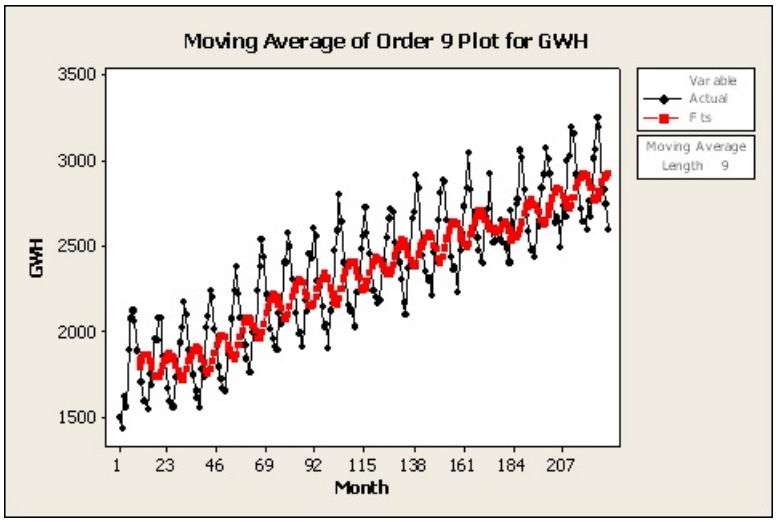
	Moving Averages of Order 12				
	1st	2nd	3rd	4th	5th
Jan-78	1503	1503	1503	1503	1503
Feb-78	1438	1438	1438	1438	1438
Mar-78	1629	1629	1629	1629	1629
Apr-78	1564	1564	1564	1564	1564
May-78	1897	1897	1897	1897	1897
Jun-78	2079	2079	2079	2079	2079
Jul-78	2131	2131	2131	2131	2131
Aug-78	2065	2065	2065	2065	2065
Sep-78	1894	1894	1894	1894	1894
Oct-78	1848	1848	1848	1848	1848
Nov-78	1713	1713	1713	1713	1713
Dec-78	1596	1596	1596	1596	1596
Jan-79	1598	1598	1598	1598	1598
Feb-79	1548	1548	1548	1548	1548
Mar-79	1760	1760	1760	1760	1760
Apr-79	1686	1686	1686	1686	1686
May-79	1964	1964	1964	1964	1964
Jun-79	1956	1956	1956	1956	1956
Jul-79	2092	2092	2092	2092	2092
Aug-79	2085	2085	2085	2085	2085
Sep-79	1858	1858	1858	1858	1858
Oct-79	1841	1841	1841	1841	1841
Nov-79	1671	1671	1671	1671	1671
Dec-79	1596	1596	1596	1596	1596

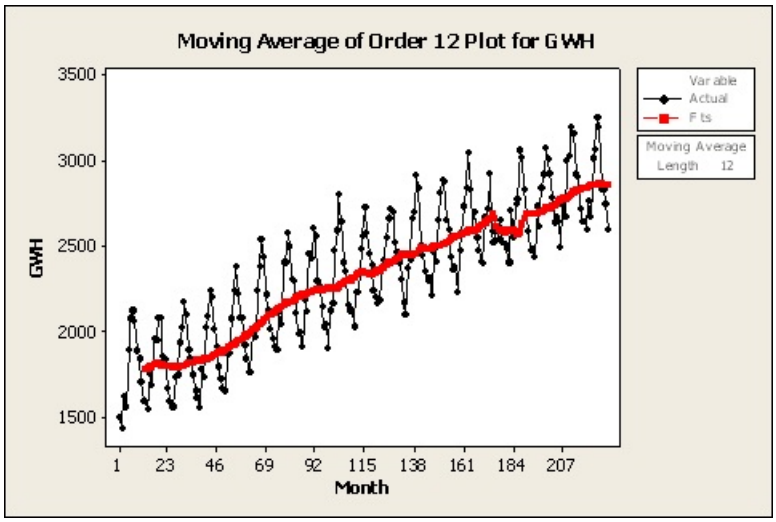
Each moving average contains one January, one February,....

Centreing:
The moving averages correspond best to the 6 ½ month, so two successive moving averages are further averaged, to make it correspond to a particular month

A Moving Average of Order = Length of Season removes Seasonal Effects and the Erratic Component, leaving Trend.







So far so good, but now we have some choices.

Y_t = actual value of series at time t.

S_t = the seasonal effect (index) at time t.

T_t = the trend at time t.

E_t = the erratic component

To simplify we usually consider only two models, shown here for monthly data:

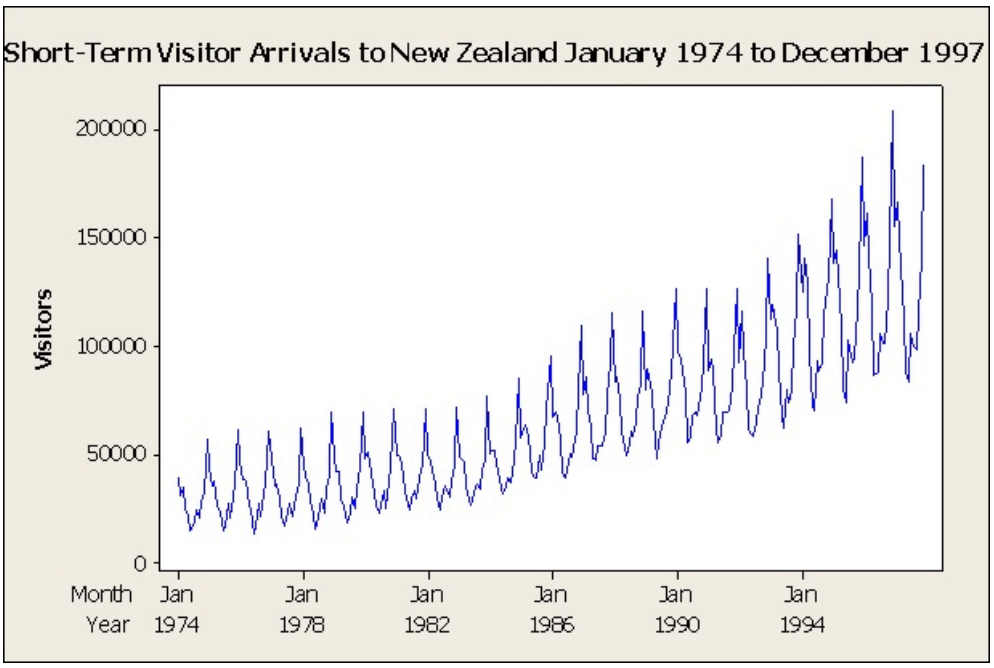
Additive Model

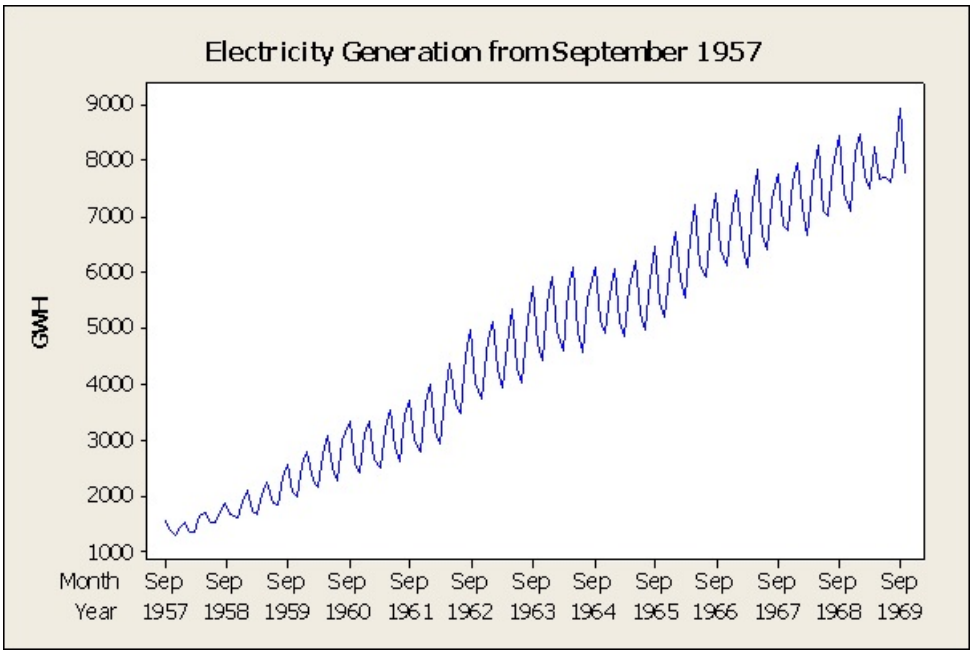
$$Y_t = T_t + S_t + E_t \qquad t = 1,2,...; \qquad \sum_{any\ year} S_t \doteq 0$$

Multiplicative Model

$$Y_t = T_t . S_t . E_t \qquad t = 1,2,...; \qquad \sum_{any\ year} S_t \doteq 12$$

Scholarship usually appears to assume an additive model, even though multiplicative (seasonality) is much more common.





The General Method (Classical Decomposition) (assuming the additive model)

- Find indices to describe the seasonal pattern:
- Subtract these seasonal indices to remove seasonal effect.
- Forecast the deseasonalised series
- Add the relevant indices to restore seasonal pattern to forecasts.

This assumes additive seasonality. When seasonality is multiplicative, replace addition and subtraction by multiplication and division.

Scholarship usually appears to assume an additive model

The Scholarship Question/Method

1. Doesn't usually specify that the series is seasonal, but it is usually expected.

2. Usually provides the centred moving averages (CMA's)

3. Deseasonalises by subtracting the CMA's

4. May require averaging the individual seasonal effects

5. Forecasts the deseasonalised data usually by fitting a linear regression line to the CMA's. Often several lines to choose from.

6. Reseasonalises (if needed) by adding in the appropriate seasonal indices.

In Symbols:

$$Y_t = T_t + S_t + E_t$$

Now CMA's remove the Seasonal and Erratic Components,
so you could say we have formed:

$$X_t = T_t$$

Subtracting the CMA's leaves the Individual Seasonal Effect:

$$S_t + E_t$$

Averaging all the Januarys, and all the Februarys,..... gets rid of the Erratic Variation to produce the Seasonal Indices

An Aside - Ignore this when studying!

What's wrong with the Scholarship Method?

Fitting a form (linear, quadratic, exponential) to the trend line is unnecessary and may encourage excessive extrapolation, as happened in electricity generation in the 1970's

The Politics of Energy Forecasting, Baumgartner&Midttun, Oxford,1987

West Germany

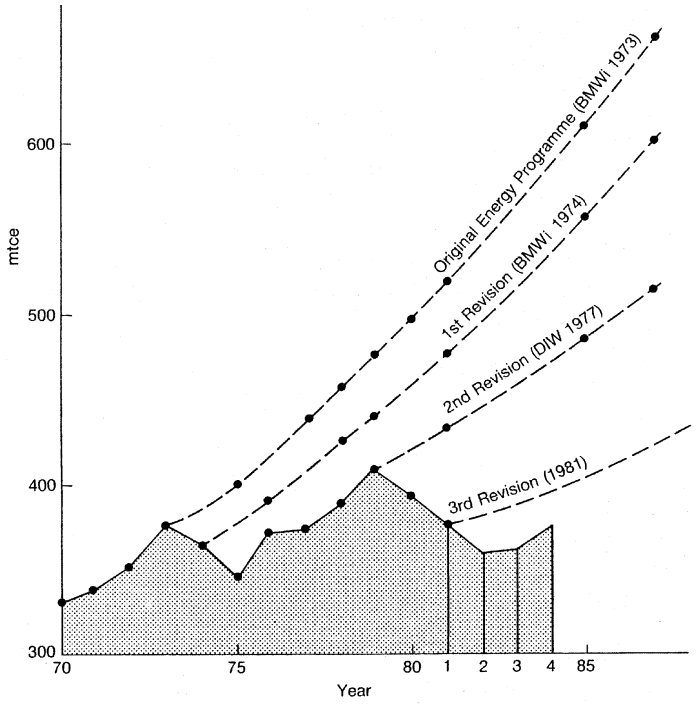


Fig. 4.4 Comparison between actual primary energy consumption and the forecasts of the Federal Energy Programmes

- It places equal weight on the oldest and the newest data.

- Regression is set up to get the answer right about the middle of the data set, not at the right end (the future).

- The last 7 months of known values (assuming monthly data) are (effectively) replaced by extrapolations!

Comparing the Last Six Years' Scholarship Questions

	AME Book	Series	Seasonal?	CMA provided?	Forecasting Method(s) Proposed	Seasonality Reintroduced into Forecasts?
2004 Q.6	**Q.2**	Monthly Daily	Strongly clearer from table - crappy Excel graphs!	Yes "CMM order 12" Yes "CMM order 7"	Straight line	Expected, using both monthly and daily indices. Heavens!
2005 Q.6	**Q.3**	Monthly	Not clear	Yes "12-point"	Straight line. Quadratic	Expected
2006 Q.6	**Q.1**	Monthly	Yes, from graph	Yes "12-point CMM"	None	Expected, and in report discussion
2007		No forecasting question. Time series on defective kiwifruit but answers cross-sectional				
2008 Q.3		Quarterly	Not clear to me	"Deseasonalised"	Straight line. Extrapolation from last 5 values of CMA's	Expected in report and forecasts
2009 Q.4		Quarterly	Yes, Dec higher than March	Yes ""CMM"	3 equations Pick one with highest R^2 and right trend	Expected

A Few Points

- There are a lot of marks to be gained by describing features of a series - trend, seasonality, degree of erratic variation - **if these are present**. (Seasonal effects must always be calendar related 5, 7, 4, 12)

- Words like “deseasonalised”, “centred moving mean” strictly do not imply there is seasonality. CMA’s can be used to remove erratic components in non-seasonal series. However almost all the questions have assumed seasonality.

- Use tables as well as graphs to diagnose. It may be hard to read the components off the graphs. Get a plastic ruler!

- While multiplicative seasonality is more realistic it has always been additive so far.

- Requiring comment on the dangers of excessive extrapolation is very common.

- A common thing is to give two potential trend lines: linear or quadratic, all data or most recent data (2004, 2005, 2008, 2009). Possible criteria: highest R^2, capturing recent trend. **Sketch the lines on the question book graph.**

- Any forecasting question may be mixed up with linear regression, linear equations. **Read this question particularly carefully**! Identify the techniques to use.

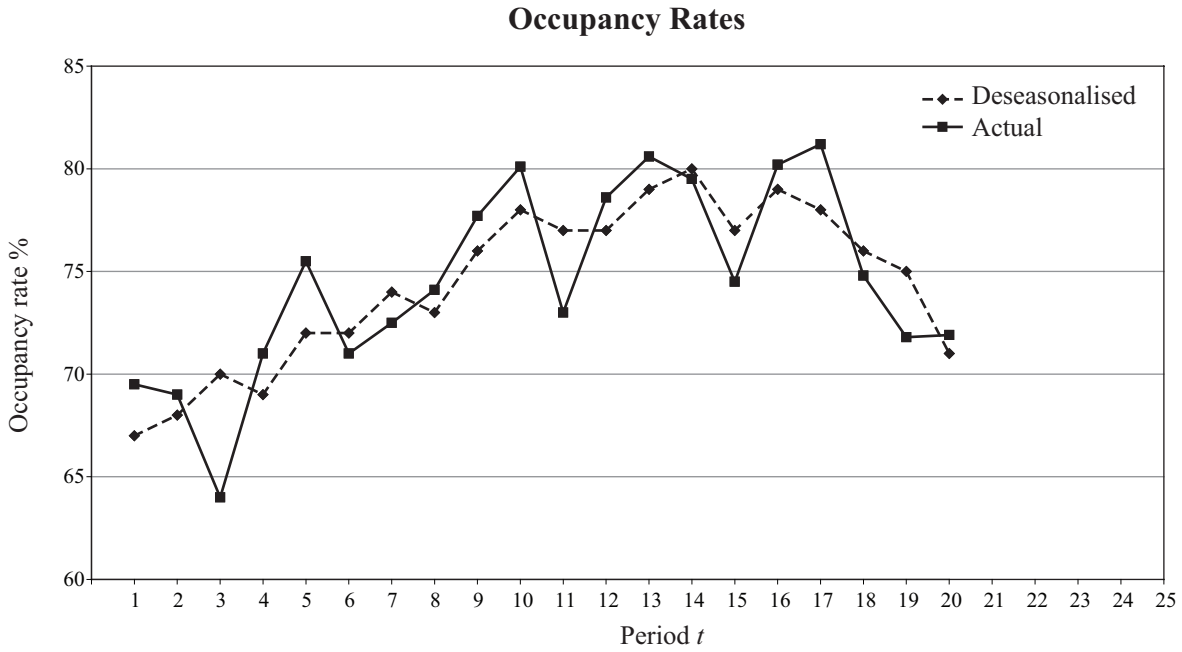
- Turn answers back into proper units.

QUESTION THREE 2008: OCCUPANCY RATE INVESTIGATION (8 marks)

Statsmod Enterprises wishes to investigate the trend in the occupancy rate of its hotel in the period 2003 to 2007. Data for the actual occupancy rate for each quarter of the years 2003 to 2007 were collected as shown in the following table. Period $t = 1$ is the first quarter in 2003, period $t = 2$ is the second quarter of 2003, and so on. Deseasonalised occupancy rates were calculated to the nearest whole number and these are also shown in the table.

Period t	Actual occupancy rate (%)	Deseasonalised occupancy rate (%)	
1	69.5	67	2.5
2	69.0	68	
3	64.0	70	
4	71.0	69	
5	75.5	72	3.5
6	71.0	72	
7	72.5	74	
8	74.1	73	
9	77.7	76	1.7
10	80.1	78	
11	73.0	77	
12	78.6	77	
13	80.6	79	1.6
14	79.5	80	
15	74.5	77	
16	80.2	79	
17	81.2	78	3.2
18	74.8	76	
19	71.8	75	
20	71.9	71	

From the data in the table the following graph was obtained.



A trend line was fitted to the deseasonalised data. Its equation is $y = 0.4496\,t + 69.679$, where y is the deseasonalised occupancy rate as a percentage.

(a) Write a short paragraph to describe occupancy rates over the years 2003 to 2007.

(b) Obtain a forecast for the occupancy rate in the first quarter of 2008 by each of the following methods:

(i) Extrapolate the deseasonalised occupancy rates (%) corresponding to $t = 16$, 17, 18, 19 and 20. Use the grid provided on page 25 of the Answer Booklet.

(ii) Use the fitted trend line.

(c) Discuss any differences between your answers in part (b).

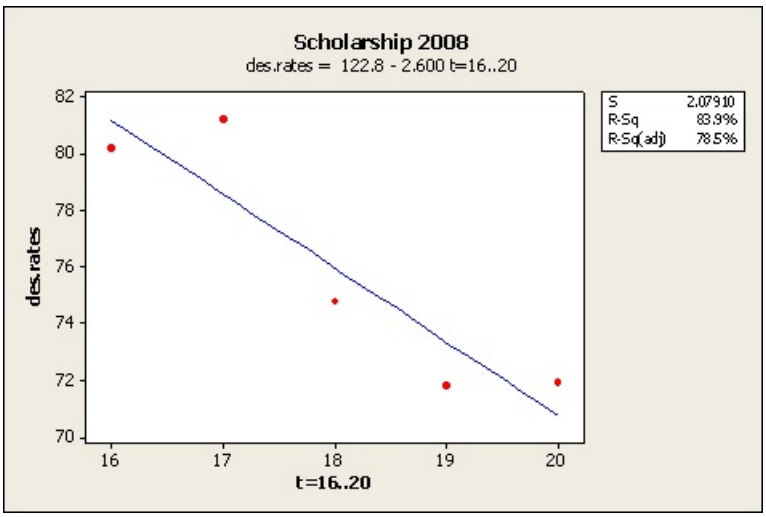
Question 3 2008

Forecast Q1 2008 (t=21)

The average of all the first quarter seasonal effects is:

$$S_1 = \frac{2.5+3.5+1.7+1.6+3.2}{5} = 2.5$$

I. From the points t =16..20.

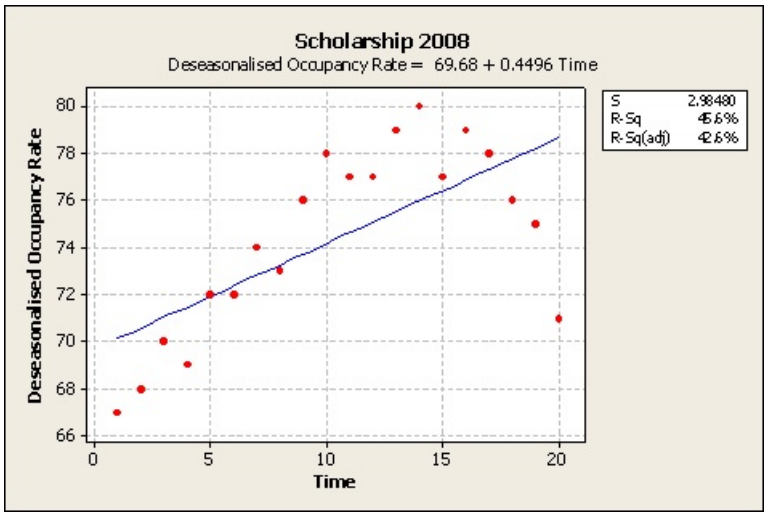


122.8 - 2.6(21) = 68.2

Forecast = 68.2+2.5 = 70.7%

Answers: "Extrapolating the deseasonalised graph we get 70% corresponding to t = 21. Can have any reasonable value in the range 68 to 72% when extrapolating"

2. From the trendline.



0.4496(21) + 69.68 = 79.1

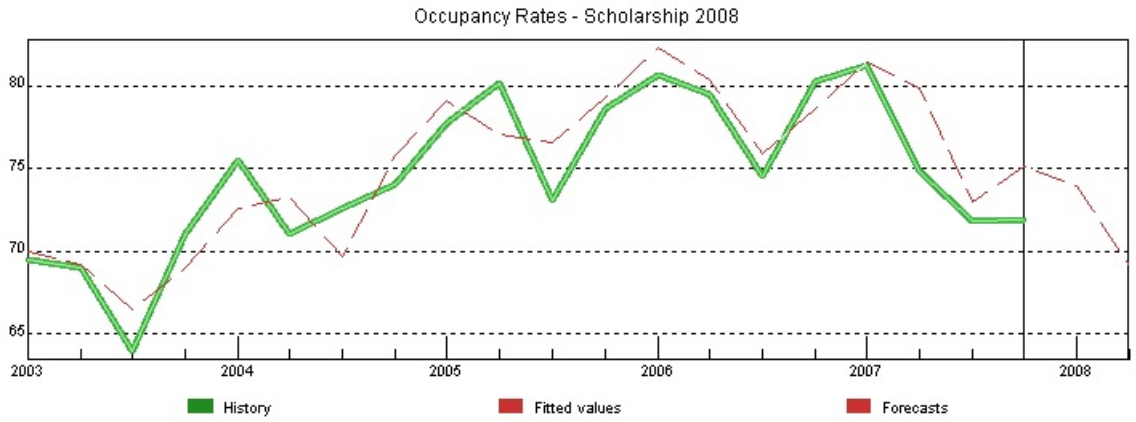
Forecast = 79.1 + 2.5 = **81.6%**

Points to make:

Recent data better. Trend appears to have changed. Full trend line gave equal weight to old data, so it doesn't pick up downwards trend in last 4 quarters. R^2 much better.

An Aside:
What's Actually Used in Forecasting Packages:

Adaptive methods, avoiding the need to specify any form for trend, or to extract seasonality. E,g ,Forecast Pro



Model Details				
Expert selection				
Additive Winters: Linear trend, Additive seasonality				
LA(0.251, 1.000, 0.269)				

Component	Smoothing Wgt		Final Value	
Level	0.2512		73.44	
Trend	1		-2.315	
Seasonal	0.269			
Seasonal Indexes				
Periods 1-4	2.827	0.2862	-3.342	0.2282

Forecast Data				
Date	**2.5 Lower**	**Forecast**	97.5 upper	
2008-Q1	68.35	73.96	79.56	
2008-Q2	62.83	69.1	75.37	

Take MSCI 202 "Forecasting and Simulation" to learn more!

Linear Regression

We want to fit an equation of the form: $Y = a + b\,X$
to data that is a set of bivariate pairs (Y_i, X_i), using the least squares of the errors (residuals) principle.

$$Y_i = a + bX_i + e_i$$

So we find values of ***a*** and ***b*** so that the sum of squares of the errors is as small as possible.

Why Least Squares?

- It's easy - a formula or solving linear equations
- Squared errors - big errors bad, small errors just "noise"
- It gives a unique line

What are the unknowns in this problem? What makes it linear?

- ***a*** and ***b***, not X and Y.
- In particular values of X may be specified e.g. time,
- or modified, e.g. $Y = a + b\,ln(X)$, *or* $Y = a + b\,X^2$

Why a linear relationship?

- Simplicity
- A localised (approximate) relationship
- We don't have any reason for a more complex form
- Variable transformations, piecewise fits, can get around the limitation of a straight line fit.

These lead to a lot of the Scholarship questions:

- Outliers
- Transformations which give better fits
- Avoiding too much extrapolation
- Piecewise fits

Where do the formulas for ***a*** and ***b*** come from?

(Definitely not in the syllabus)

The value for Y_i predicted by the equation is:

$$\hat{Y}_i = a + bX_i$$

So:

$$\sum_{i=1}^{n} (e_i)^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$$

$$= \sum_{i=1}^{n} (Y_i - a - bX_i)^2$$

To minimise this:

- differentiate with respect to ***a*** and ***b***,
- put the two partial derivatives equal to zero
- and solve the resulting two equations

Let's not go any further down this track!

QUESTION THREE

A market research company is contracted by the general manager of NAILS to investigate the effect of promotional expenditure on sales of plants. The results of the investigation will be used to make sales predictions for all gardening departments in the New Zealand-wide store chain.

For each month from January 2001 to December 2002, the amount spent on promoting plant sales (promotional expenditure) and the total plant sales by the chain were recorded. The data, a scatter plot and some statistical output relating to the data are shown below. E represents the promotional expenditure for the month (in thousands of dollars, \$000) and S represents the total plant sales for the month (in thousands of dollars).

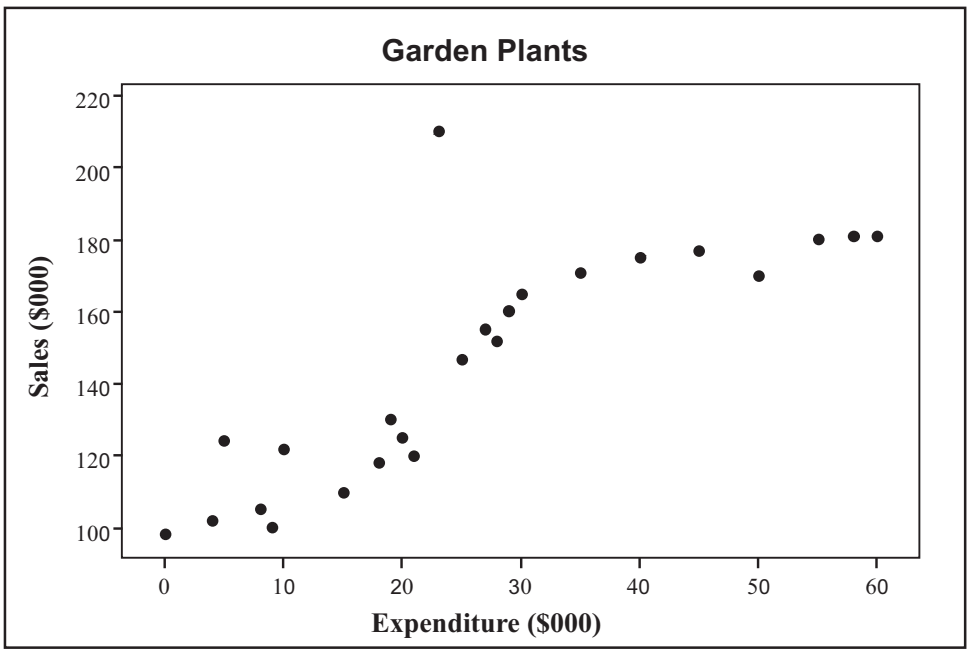
Write a report (approximately one page long) to the general manager of NAILS that summarises the statistical output given below. Include some sales predictions in your report.

Data Output for Question Three

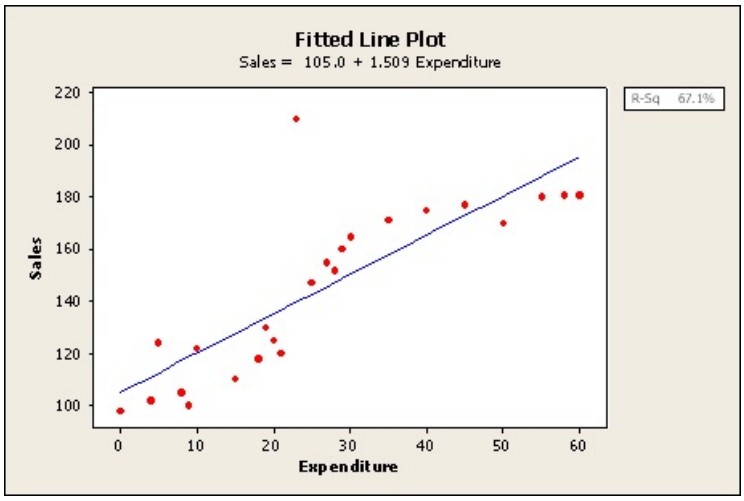
Expenditure (\$000)	Sales (\$000)
0	98
4	102
5	124
8	105
9	100
10	122
15	110
18	118
19	130
20	125
21	120
23	210
25	147
27	155
28	152
29	160
30	165
35	171
40	175
45	177
50	170
55	180
58	181
60	181

Value of E	Correlation r	Regression Equation
All values	0.819	$S = 105 + 1.51E$
All less outlier	0.928	$S = 101 + 1.54E$
$E \geq 15$ less outlier	0.885	$S = 104 + 1.47E$
$15 \leq E \leq 30$ less outlier	0.972	$S = 53.2 + 3.67E$

Scatter Plot



Question 3, 2004

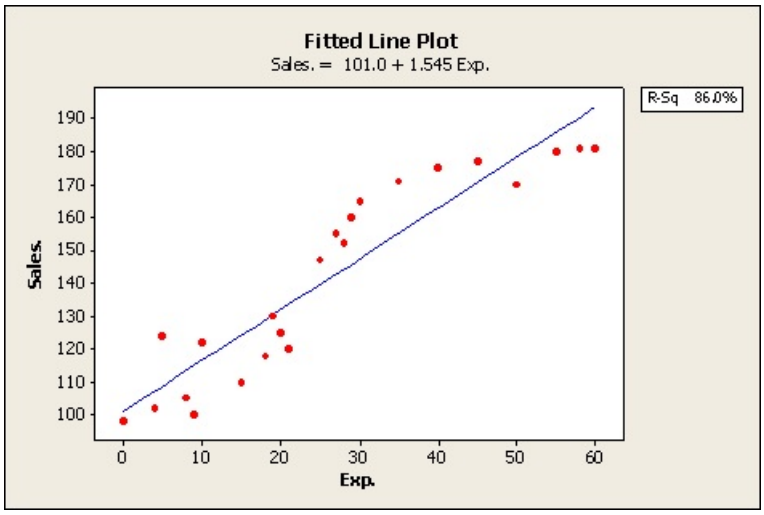


A bit of the output:

```
The regression equation is
Sales = 105 + 1.51
Expenditure


Unusual Observations

Obs  Expenditure   Sales      Fit  SE Fit  Residual  St Resid
 12         23.0  210.00   139.76    3.94     70.24      3.80R

R denotes an observation with a large standardized residual.
```

Observation 12 is having a large effect on the line. Ideally we should be able to confirm that it is an error. After deleting it:

```
The regression equation is
Sales = 101.0 + 1.545 Exp.


R-Sq = 86.0%
```



Could argue for 3 regions. Moderate positive correlation, outlier removal good, use the piecewise linear approximation given, weak positive correlation up to $15,000, moderate positive correlation in range $15-30,000, saturation at sales = $180,000? Do a prediction within the range of the best regression - e.g. expenditure = $25,000

We stop here to answer the question: why does Minitab refer to R^2 as a percentage?

The Coefficient of Determination, R^2

(This is background - don't panic about it)

Exp.	Sales.	Fitted Values	Residuals
0	98	101.046	-3.0462
4	102	107.226	-5.2259
5	124	108.771	15.2292
8	105	113.405	-8.4055
9	100	114.950	-14.9504
10	122	116.495	5.5047
15	110	124.220	-14.2198
18	118	128.855	-10.8545

.....

Correlations: Exp., Sales., Fitted Values

	Exp.	Sales.
Sales.	0.928	
Fitted Values	1.000	0.928

It's the square of the correlation between *Y* and *X*:

$$.928^2 = 0.86$$

It's the square of the correlation between *Y* and the Fitted Values, $\hat{Y}$

We can write:

$$\sum_{i=1}^{n}(Y_i-\bar{Y})^2 = \sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 + \sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$$

i.e. Total = Regression + "Residual Error",

which Minitab handily spits out:

Analysis of Variance

Source	Sum of Squares
Regression	16764
Residual Error	2718
Total	19482

Thus another interpretation of R^2 is that it is the fraction (or percentage) of variation "explained" by the regression line, i.e.:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}{\sum_{i=1}^{n}(Y_i-\bar{Y})^2} = \frac{16764}{19482} = 0.86$$

e.g. Suppose the regression line fitted the data perfectly. Then the residual sum of squares is zero, i.e. $R^2 = 1$ (100%).

Hence the name: **Coefficient of Determination.**

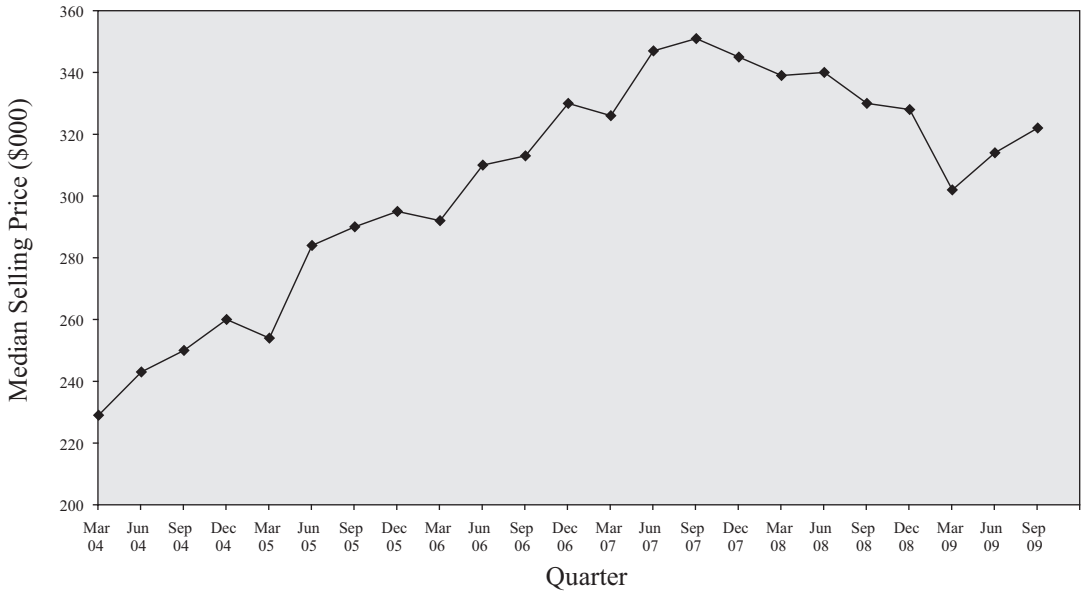
QUESTION FOUR 2009 (8 marks)

In an investigation of home prices, the median home selling price, in thousands of dollars ($000), was obtained for each quarter from March 2004 to September 2009. Centred moving means (rounded to the nearest thousand dollars) were calculated. The data are shown in Table 3 below, along with an inflation index (Base: June quarter 2005 = 1000). The median home selling prices are displayed in Table 3 and a graph of median home selling prices by quarter is displayed in Figure 3.

Quarter	Median Home Selling Price ($000)	Centred Moving Mean ($000)		Inflation Index
Mar 04	229			
Jun 04	**243 x1000/284**			924
Sep 04	250	249		
Dec 04	260	257		
Mar 05	254	267	-13	
Jun 05	**284 x1000/284**	276		1000
Sep 05	290	286		
Dec 05	295	294		
Mar 06	292	300	- 8	
Jun 06	**310 x1000/284**	307		1047
Sep 06	313	316		
Dec 06	330	324		
Mar 07	326	334	-8	
Jun 07	**347 x1000/284**	340		1208
Sep 07	351	344		
Dec 07	345	345		
Mar 08	339	341	-2	
Jun 08	**340 x1000/284**	336		1224
Sep 08	330	330		
Dec 08	328	322		
Mar 09	302	318	-16	
Jun 09	**314 x1000/284**			1249
Sep 09	322			

Table 3

Figure 3: Median Home Selling Prices, March 2004 to September 2009



Several lines (given below) were fitted to the data and the following equations were obtained, where y is the median selling price ($000) and x is the number of quarters since December 2003 (ie, $x = 1$ for the March 2004 quarter, $x = 2$ for the June 2004 quarter, etc).

Equation 1: $y = 4.27x + 253$ with $R^2 = 0.64$ fitted to the median sale prices.

Equation 2: $y = 4.63x + 254$ with $R^2 = 0.72$ fitted to the centred moving means.

Equation 3: $y = -5.66x + 437$ with $R^2 = 0.99$ fitted to the centred moving means for the period December 2007 quarter to March 2009 quarter.

(a) Write a short paragraph to describe median home selling prices from 2004 to 2009.

(b) Calculate a forecast for the median home selling price for the March quarter 2010. Show full working and justify your method. Discuss the validity of your forecast.

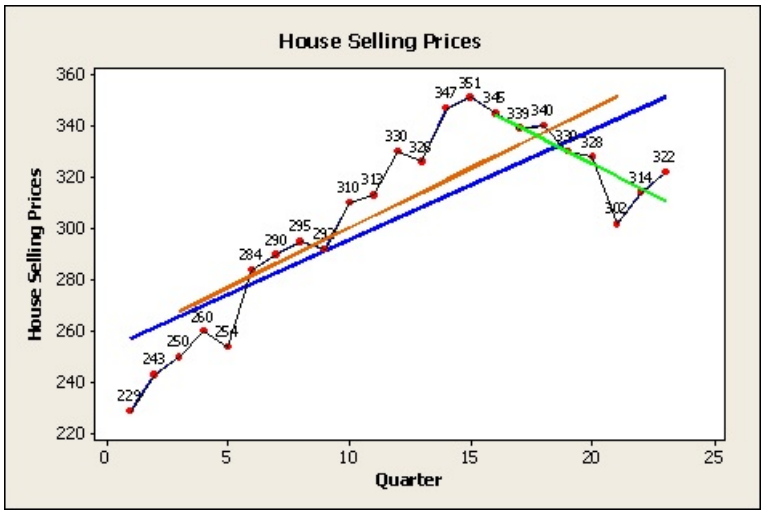
(c) Develop an index number series for the median home selling price that gives a value for the June quarter each year with the same base as the inflation index. Comment on the change in median home selling prices compared with inflation in the June quarters over the years 2004 to 2009.

i.e. median hone selling price is not to include inflation, nothing to do with centred moving mean

~~**Questions Five and Six are on the following page.**~~

Question 4, 2009

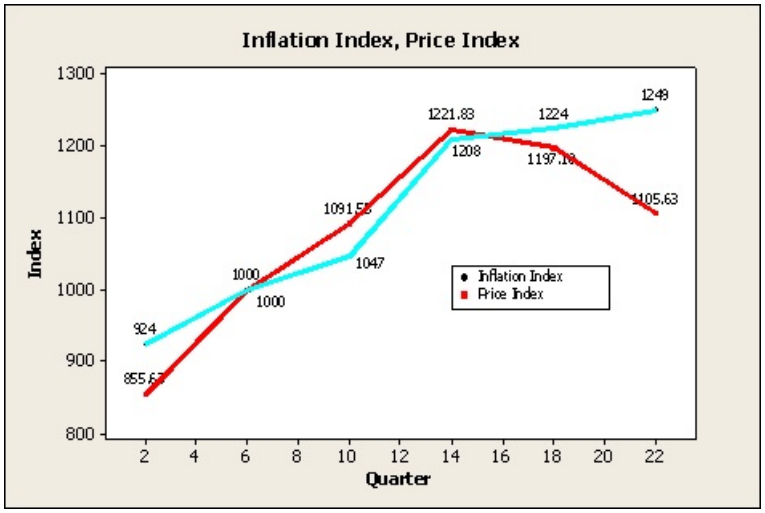
(b) *Look at the graph, don't extrapolate from very old data, need negative slope.*
Trend appears to have changed at about Sept 2007, so equation 3 is the only possibility (also has R^2 value close to 1)



For March 2010, x = 25.
Deseasonalised forecast = 437 - 5.66(25) = 295.5
For March quarter, seasonal index is -9.4, so
Forecast = 295.5 – 9.4 = 286.1 = $286,100
Trend could have changed in June and September 09, so forecast not very certain.

(c) Take June Price values, divide by 284 and multiply by 1000.

Between June 2004 and June 2007, the median home prices have increased at a faster rate than inflation.



In June 2008 and June 2009, the median home prices fell in real terms (slower rate than inflation)

An Extension: (definitely not in the syllabus)

The idea of linear regression extends easily to several explanatory variables in the same equation.

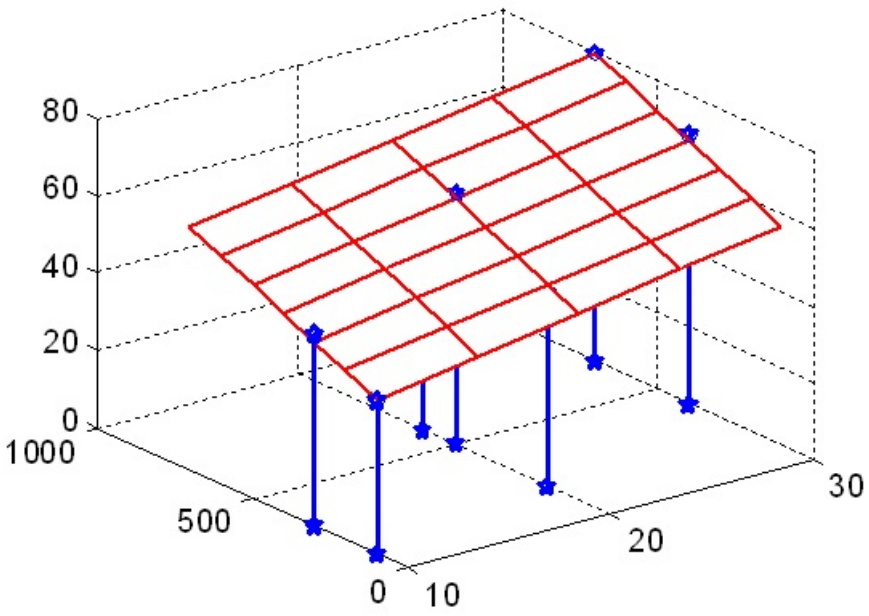
We have applied different amounts of fertiliser to 7 fields that got different amounts of rain, and measured the yield

Yield(tonnes)	Fertilizer(kg)	Rainfall(inches)
40	100	10
50	200	20
50	300	10
70	400	30
65	500	20
65	600	20
80	700	30

Regression Analysis: Yield versus Fertilizer, Rainfall

```
The regression equation is
Yield = 28.1 + 0.0381 Fertilizer + 0.833 Rainfall

R-Sq = 98.1%
```

We are now fitting a plane to the data instead of a line, but still using the least-squares method.



- The extension to two or more explanatory variables is very easy

- We can have “dummy” - variables (variables that take only the values zero or one) to account for group membership. E.g. gender, code as male = 0, female = 1

- The regression coefficients have a nice marginal interpretation:

```
The regression equation is
Yield = 28.1 + 0.0381 Fertilizer + 0.833 Rainfall
```

Other things being equal, the effect of applying 1 extra kg of fertilizer is to increase the yield by .0381 tonnes.

Other things being equal, the effect of 1 extra inch of rain is to increase the yield by .833 tonnes

Take MSCI 280 “Statistical Methods for Management” to learn more!